From Pictorial Structures to Deformable Structures

Silvia Zuffi[1,4] Oren Freifeld[2] Michael J. Black[1,3]
[1]Department of Computer Science and [2]Division of Applied Mathematics,
Brown University, Providence, RI 02912, USA
[3]Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany
[4]ITC - Consiglio Nazionale delle Ricerche, Milan, Italy

Abstract

Pictorial Structures (PS) define a probabilistic model of 2D articulated objects in images. Typical PS models assume an object can be represented by a set of rigid *parts connected with pairwise constraints that define the prior probability of part configurations. These models are widely used to represent non-rigid articulated objects such as humans and animals despite the fact that such objects have parts that deform non-rigidly. Here we define a new Deformable Structures (DS) model that is a natural extension of previous PS models and that captures the non-rigid shape deformation of the parts. Each part in a DS model is represented by a low-dimensional shape deformation space and pairwise potentials between parts capture how the shape varies with pose and the shape of neighboring parts. A key advantage of such a model is that it more accurately models object boundaries. This enables image likelihood models that are more discriminative than previous PS likelihoods. This likelihood is learned using training imagery annotated using a DS "puppet." We focus on a human DS model learned from 2D projections of a realistic 3D human body model and use it to infer human poses in images using a form of non-parametric belief propagation.*

1. Introduction

Pictorial Structures (PS) represent objects as a collection of rigid parts that can be connected in a range of spatial relationships [14]. The variability in the spatial configuration of the parts enables such models to represent variability within an object class; for example, they can capture the articulated structure of the human body in which parts are related to each other by relative rotations. In current uses such models map conveniently to a probabilistic graphical model formulation that combines image observations with pairwise potential functions encoding spatial relationships. Consequently, PS models are in wide use [3, 9, 13, 27, 29, 30].

A key limitation of PS models is that the parts themselves are treated as rigid templates (with some exceptions to allow foreshortening, *e.g.* [32]). This limits the expressive power of the model. While PS models are often called *deformable*, the deformation is in terms of the relative spatial arrangement of the parts. In general, however, the parts of natural articulated objects vary in shape *as a function of their spatial arrangement*. For example, rotation of a human upper arm causes changes in shoulder shape.

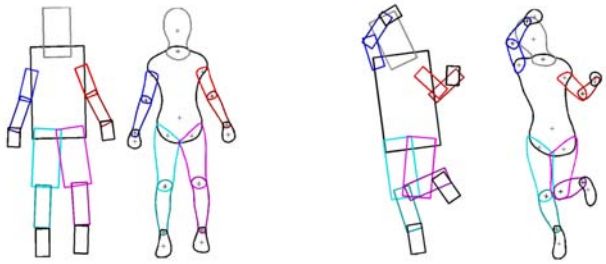

Figure 1. Deformable structures (right model in each pair) are similar to PS models but capture 2D body shape deformations.

Here we define a new Deformable Structures (DS) model as illustrated in Figure 1. The DS model has several key differences from standard PS models. First, each part is represented by a deformable contour. The shape of the contour lies in a low-dimensional linear subspace learned using Principal Component Analysis. The model is learned from a realistic part-based 3D body model that is projected into the image, producing 2D part contours. Second, the pairwise potentials between parts capture how the part shapes vary. The shape of a part depends on the shape of its neighboring parts and the relative angles between them. We model these variations with simple linear Gaussian models; while simple, this works well in practice.

Typical PS models [3] use a likelihood term learned from labeled training data. This data does not specifically define the bounding contour of the person and the shape variation of the person is accounted for *implicitly* in the learned likelihood model. Our approach is quite different. By making body shape explicit, we reduce the work required of the

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

likelihood, enabling it to be more precise. We then train a likelihood model using similar features to previous work but with features defined relative to the location of the part contour. Our approach is in line with recent work on building structured, longer-range, descriptors of 2D human appearance [2, 5, 30, 35, 36].

Our focus is on defining the DS model representing natural human body shapes. We systematically evaluate its performance in 2D human pose estimation on the *Buffy the Vampire Slayer* data set [13]. Finally, while we develop the model in the context of 2D human body pose and shape, the model itself is fully general and can be used to represent other articulated objects with non-rigidly deforming parts.

2. Background and Related Work

Likelihoods versus priors. PS models are effective for detecting people in images in a variety of poses [3, 9, 13, 27, 29, 30]. The rectangular parts in current methods however, do not capture the shape of real body parts. Rather, PS models provide a prior probability distribution over the articulated structure of the body. In a Bayesian framework, this leaves the likelihood model to capture all the non-rigid structure of the limbs. A good likelihood model has proven critical for good PS performance.

Most accurate PS methods learn a likelihood function as the normalized score of a classifier trained on rectangular boxes containing body parts. The features used for the classifier can vary and include shape contexts [3], histograms of orientated gradients (HOG) [8, 30], and raw image pixels [28]. Boosted classifiers or SVMs are then typically used for learning. Despite advances, current likelihoods are actually quite bad at body part detection [24].

Many recent methods address the problem of learning and using richer likelihoods. Sapp *et al*. [29] use cues based on shape similarity defined by regions and contours. Another approach learns discriminative appearance models that depend on pose by first clustering the space of possible poses [18, 20, 37]. To train such models, Johnson *et al*. use Amazon Mechanical Turk to collect a large data set of annotations [21]. Alternatively, 3D body models with varying shape and pose can be used to generate synthetic training images [25]. Wang *et al*. [38] argue that the anatomical concept of a "part" is a limiting construct of part based models. Instead, they propose a hierarchical representation based on Poselets [5] where parts can include more than one anatomical limb. Another approach defines *part type* as an indicator of semantic and geometric attributes (*e.g*. a "stretched" or "fore-shortened" arm) enabling the sharing of example parts of similar shape across different object poses [35]. More complex likelihoods, however, come with a computational cost. To deal with this, Sapp *et al*. [30] use a cascade approach to sequentially prune the state space while using increasingly complex models.

Our work is similar in spirit to the above but takes a different approach. We use a body shape model to capture the *predictable variability* in body shape. The image likelihood term then is required to do less; it can focus on modeling how the body shape model relates to image measurements rather than modeling body shape itself.

The potential of potentials. The PS model is a tree-structured graphical model with parts represented by nodes in the graph and the spatial relationships between parts represented by potential functions [12, 14]. These potential functions typically define a probability distribution over part-joint locations (*e.g*. like a spring) and the relative angles between parts. Felzenszwalb and Huttenlocher [12] also define a scale factor for each part to account for fore-shortening and a Gaussian potential over the difference of scale between neighboring parts. A Gaussian formulation of the potentials admits efficient inference [11]. We go beyond scaling the parts and allow them to have a range of deformations learned from training examples. Like previous methods the potentials are Gaussian but are defined jointly over pose and shape parameters. The distribution over the shape, location, and pose of a part is conditioned on the shape, pose, and location of its neighbors.

Other 2D models. Active Shape Models [7] represent shape with eigenvectors of the covariance matrix of the contour points in a training set of aligned shapes. Our model of parts is similar but we go further to define how part shapes change with articulation.

Several methods go beyond PS and tie pose estimation to segmentation, making a model of 2D body shape important. ObjCut is a generative deformable object-specific MRF model that combines elements of PS models with spatial MRFs [22]. PoseCut [6] optimizes 3D body pose to best segment an image but relies on a crude 2D shape model. Predicting body shape from 2D poses enables Wang and Koller [36] to improve pose estimation.

The DS model is similar to the Contour Person (CP) [15] which is also a 2D model of body shape learned from 3D models of people in different poses. CP, however, is quite different in that it is a "global" model. While CP "factors" camera view, shape, and pose, it is not factored in the sense of a PS model. DS and PS provide a Bayesian factorization of the posterior probability, admitting tree-based inference algorithms. In DS, the local part-based shape representation allows independent rendering of body parts contours, supporting local evaluation of part-based likelihoods.

3. Model definition

The basic PS model. Following [12], let Θ denote the parameters of the model, I the image data, and L a configuration of the object, namely the location and orientation for each object's part. The posterior distribution characterizing the probability of the object configuration given the model

and the image can be expressed following Bayes' rule as

$$p(L|I,\Theta) \propto p(I|L,\Theta)p(L|\Theta) \quad (1)$$

where $p(I|L,\Theta)$ is the image likelihood given the model configuration and its parameters, and $p(L|\Theta)$ is the prior probability of the configuration of the model, given the model parameters. For pictorial structures the likelihood term is approximated as the product of the likelihoods of the individual parts. This assumes conditionally independence of part likelihoods, which is not true when parts can overlap but simplifies the inference problem. With this factored likelihood we rewrite (1) as

$$p(L|I,\Theta) = \frac{1}{Z} \prod_{i=1..M} \phi_i(\mathbf{l}_i) \prod_{(i,j)\in E} \psi_{ij}(\mathbf{l}_i, \mathbf{l}_j|\Theta_{ij}) \quad (2)$$

where M is the number of parts, E is the set of pairs of connected parts, Z is the partition function, $\mathbf{l}_i$ is the configuration of part i, $\phi_i(\mathbf{l}_i)$ are the unary potentials that include the likelihood of part i and any other part-specific prior information, and $\psi_{ij}(\mathbf{l}_i, \mathbf{l}_j|\Theta_{ij})$ are the pairwise potentials with parameters Θ_{ij}.

The potentials, $\psi_{ij}(\mathbf{l}_i, \mathbf{l}_j|\Theta_{ij})$, take the form of unnormalized Gaussian distributions over the difference between part joint locations; this is analogous to having "springs" connecting parts [14]. The basic potentials also define a probability distribution over the angle between parts.

Training data. The DS model keeps the PS representation of pose and adds parameters that provide a more realistic generative model for 2D body shape. The DS model is learned from training contours derived from SCAPE [4], a parametric 3D model of articulated human shape. Using an approach similar to [15] we generate random SCAPE poses from random cameras within a range of poses and viewing directions and project these into the image plane to create 2D training contours. We generate separate training samples for males and females using a single body shape for each; each DS model is learned from 3000 mirrored samples. Figure 2 shows example poses in the training set for the female body in the frontal view; note the variability of pose and orientation of the body relative to the camera.

Each part is rendered as a separate 2D closed contour and discretized into a fixed number of contour points plus two additional "joint" locations at the proximal and distal ends of the part. The two "joints" define a local coordinate system for the part and the line through them divides the part into two sides. Each side of the part is sampled to a fix number of points, evenly spaced according to the arclength.

Below we experiment with models having different numbers of body parts: 10 parts, consisting of the head, torso, upper and lower limbs, where the hands and feet are included in the lower limbs; and 14 parts with hands and feet treated as independent parts. We use the 10-part model during inference for a more direct comparison to traditional PS models. Additionally we can learn separate DS models from different viewpoints (*e.g.* from the side); we do not pursue this here.

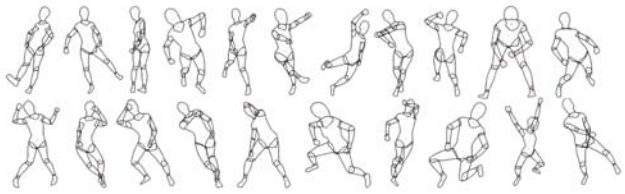

Figure 2. Examples of training poses. Note the variability in pose as well as in camera location. The red dots correspond to joint locations or other control points (zoom for detail).

Part deformations. The shape of each part is learned independently and then these shapes are coupled in the graphical model with pairwise potentials. The training examples for each part are aligned to a common coordinate system and we vectorize the set of contour points and joint points to form training vectors. The variability in part shape is modeled using Principal Component Analysis (PCA). Specifically, we learn a low-dimensional linear model

$$\begin{bmatrix} \mathbf{s}_i \\ \mathbf{p}_i \end{bmatrix} = \mathbf{B}_i \mathbf{z}_i + \mathbf{m}_i \quad (3)$$

where $\mathbf{s}_i$ is a vector of contour points and $\mathbf{p}_i$ is a vector of joint points. The vector $\mathbf{m}_i$ represents the mean contour (and joints) of part i. $\mathbf{B}_i$ is a matrix containing the eigenvectors of the training data corresponding to the dominant eigenvalues. Finally, $\mathbf{z}_i$ is a vector of linear shape coefficients that are used to represent different part shapes.

Figure 3 shows the mean contour and joint points for the head, upper leg and torso, together with contours and joint points at 2 standard deviations from the mean, for the first three PCA basis components. Most of the joint points correspond to the centers of rotation for the parts. The extremities (head, hands and feet) have one distal point that is not an anatomical center of rotation. The torso has 6 joint points: shoulders, hips, neck and belly button. While not an anatomical joint, the belly button is used to compute the part orientation and length as described below. The first component of the PCA representation typically corresponds approximately to foreshortening along the major axis. For the lower arms and legs (10-part model) the first component also accounts for some rotation of the hand or foot.

The basic deformable structures model follows the PS formulation and has the same basic parts (Figure 3). However we extend the state space for each part to be

$$\mathbf{l}_i = (\mathbf{c}_i, \sin(\theta_i), \cos(\theta_i), \mathbf{z}_i) \quad (4)$$

where $\mathbf{c}_i$ represents the location of the part center and θ_i is its orientation. Unique to the DS model are the *shape parameters*, $\mathbf{z}_i$, which define the part shape and the location of the part joints.

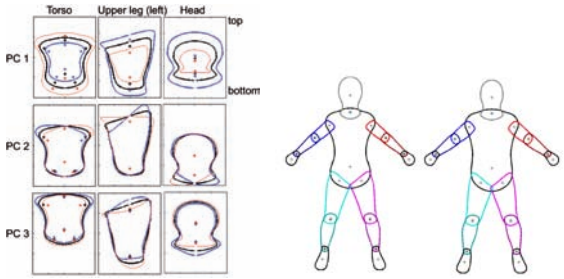

Figure 3. DS part deformations. *(left)* Deformations for three example parts. Black is the mean contour. Red and blue are ± 2 standard deviations from the mean along the first 3 principal component directions. Stars mark the joint locations which deform with the contour. *(right)* Mean part shapes for the female and male body (14-part model). The dots represent joint points (see text).

Pairwise potentials. The DS pairwise potentials relate the shape coefficients of a part to the shapes and relative orientations of neighboring parts. While these relationships could be quite complex, we find that a reasonable model is obtained with a simple Gaussian model.

Let i and j be two connected parts. The pairwise model between part i and part j is a multivariate Gaussian $\psi_{ij}(\mathbf{l}_i, \mathbf{l}_j | \Theta_{ij}) =$

$$\mathcal{N}(\mathbf{z}_j, \sin(\theta_{ji}), \cos(\theta_{ji}), \mathbf{q}_{ji}, t_j, \mathbf{z}_i, t_i | \mu_{ij}, \Sigma_{ij}) \quad (5)$$

where θ_{ji} is the relative angle of j with respect to i. The vector $\mathbf{q}_{ji}$ defines the distance between the joints of the parts; that is, $\mathbf{q}_{ji} = (\mathbf{p}_{ji} - \mathbf{p}_{ij})$, where $\mathbf{p}_{ji}$ is the joint point of part j connecting j with part i and $\mathbf{p}_{ij}$ is the joint point of part i. The points $\mathbf{p}_{ji}$ and $\mathbf{p}_{ij}$ are both defined in the local coordinate system of the part i, which has its origin $\mathbf{c}_i$ at the midpoint between the joint points, and is aligned with the main axis of the part. Note that the vector $\mathbf{q}_{ji}$ is analogous to the *spring* that connects two parts in the PS model representation. The scalars t_i and t_j are the lengths of the two parts, defined as the distance between the part joints. For the torso, the part length is defined by the distance between the neck joint and the belly button. Finally, the mean and covariance $\Theta_{ij} = (\mu_{ij}, \Sigma_{ij})$ of the Gaussian model are learned using the training samples described above.

The DS model is unique in that it is a distributed representation of body shape. The assumption is that the shape of an individual body part predicts something about the location and shape of parts that share a joint with it. Figure 4 illustrates the learned model by showing samples from it. Given a part shape we generate samples from the pairwise model for the part neighbors outwards along the tree. Note how the shape of the torso defines a distribution for the orientation of the upper arms.

Figure 4 (left) shows two different torso shapes that are used as starting points for sampling from the model. Note that the sampled poses are very different from these different starting torso shapes. This is due to the fact that torso shape is very much related to body pose. In contrast, Figure 4 (right) shows sampled poses starting from differently shaped upper arms; a single arm shape does not say nearly as much about the overall pose and shape of the body. Taken together however, the collection of body parts and their spatial relationships say a good deal about body shape. Examples of various posed models are shown in Figure 5. These provide a fairly realistic representation of 2D body shape.

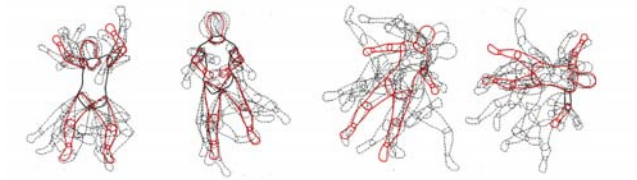

Figure 4. Sampling from the DS model. *(left)* Two different torso shapes are outlined in black. Samples from the DS model are shown as dotted black lines. These are generated by starting with the torso and moving out along the tree structure. The red contour shows the most likely pose and shape for the parts. *(right)* Two more examples starting from different shapes of the upper arm (the model is rendered in the coordinate system of the arm).

Figure 5. Examples of the DS model in a variety of poses. Note how much the model's left calf (magenta) varies in shape.

4. The DS "puppet"

There are several tools for annotating human pose in images but most give fairly crude descriptions of the body in terms of "sticks" [13]. Bourdev and Malik [5] annotate images of people with joint locations, infer a 3D body pose, and label super pixels as corresponding to different body parts or clothing. We exploit the DS shape model to provide a new annotation tool that is easy to use and directly manipulates the 2D body shape.

The interface allows a user to selectively move or lock the joint points described in the previous section. The shape of the model is inferred conditioned on these fixed points. The user sees the model deforming as he or she moves the points and can thus position it over an image.

We collected an annotated data set of 217 images. Example annotations are shown in Figure 6. These annotations are used for training the likelihood model in the following section.

Figure 6. DS puppets. The annotation tool has a Web interface for posing a draggable DS puppet over an image.

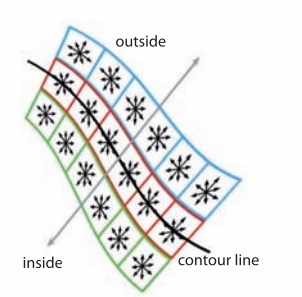

Figure 7. Contour likelihood. The image shows the location of the HOG cells along a limb contour. Cells are located on the boundary, just inside, and just outside.

5. Likelihood

The unary potentials, $\phi_i(\mathbf{l}_i)$, represent the probability of a part in a specific location in the image. Since DS defines contour points for each body part, we are able to focus the likelihood computation on the part boundary. Additionally, since we know inside from outside, it is straightforward to formulate likelihood models of skin color or textural appearance. We define the unary potentials as:

$$\phi_i(\mathbf{l}_i) = \phi_i^{\text{contour}}(\mathbf{l}_i)\phi_i^{\text{color}}(\mathbf{l}_i). \quad (6)$$

The contour based likelihood is given by

$$\phi_i^{\text{contour}}(\mathbf{l}_i) = \frac{1}{1+\exp(a_i f_i(h_i(\mathbf{l}_i)) + b_i))} \quad (7)$$

where $f_i(h_i(\mathbf{l}_i))$ is the output of a linear SVM classifier applied to the feature vector $h_i(\mathbf{l}_i)$, and a_i and b_i are scalar parameters [26].

The feature vector consists of a set of HOG descriptors computed along the part contour (cf. [23]). For a part i, we take a set of points at fixed locations in the contour vector $\mathbf{s}_i$ (3) where we compute three HOG descriptors: one at the contour point, one inside and one outside the part contour (Figure 7). All the gradients recorded by the descriptors are steered according to the local contour orientation [31]. The HOG descriptors are computed with an adaptive cell dimension, which is set on the basis of the size of the puppet in the annotated training image.

The color likelihood assumes that the lower arms are likely to be skin colored and the upper arms are likely to have the same colors as the upper torso. The color-based probability of a limb is then defined in terms of the color probability of its pixels. The probability of a pixel being skin is represented by a histogram of skin colors computed from a publicly available data set[1] and from the head regions of our training set. The probability of a pixel having the color of the upper torso is image specific, and is computed using the histogram of the pixels covered by the upper region of the torso and the head once an initial torso estimation has been provided by the inference algorithm.

6. Inference: Pose and Shape

To use the DS model for 2D pose estimation we must extend traditional PS inference to include the additional shape parameters. Like PS models, the factored form of the DS model means that inference can be done using Belief Propagation. Unfortunately, efficient BP algorithms assume a discrete (or discretized) space. When the state space of a variable cannot be enumerated, and the potentials do not allow the computation of messages in closed form, sampling approaches can be used [17, 19, 34]. Such non-parametric methods have been applied successfully in human pose estimation in 2D [32] and 3D [33]. Here we adopt a method inspired by Particle Belief Propagation (PBP) [17].

In the DS graphical model, the node variables (4) are in part discrete (joint locations in global coordinates) and continuous (angle and shape parameters). Given the large number of variables (we use 4 shape parameters per part), we formulate our inference problem as one of selecting the best configuration among discrete sets of part samples. This entails defining samples at each node, and using Max-Product BP to derive the most likely configuration. The message from node i to node j is defined over the discrete sets of N node samples, and takes the form:

$$\hat{m}_{ij}^{(q)} = \max_{p=1..N}[\psi_{ij}(\mathbf{l}_i^{(p)}, \mathbf{l}_j^{(q)})\phi_i(\mathbf{l}_i^{(p)}) \prod_{u\in\Gamma(i)\setminus j} \hat{m}_{ui}\mathbf{l}_i^{(p)}] \quad (8)$$

[1] http://acouchis.helmholtz-muenchen.de/staff/giovani/colour/

where p and q are sample indexes at node i and j respectively.

The key to obtaining good results with non-parametric BP in high dimensions is to have good proposal functions. We therefore rely on an update phase, inspired by PBP [17], where we resample parts at each node based on the current state of the neighbors' nodes. In order to provide the inference algorithm with good initial samples, we first run a person detector [13] to obtain an estimate of the scale of the person in the image. The scale is then fixed, but the DS model has been learned with variability in the camera location, and can represent some size variation in the parts. Additionally, to provide a good initialization for the part location and orientation, we run a standard PS inference method [3] as a pre-processing stage. Running a simple PS model to prune the search space is a common strategy [30, 36]. We generate part samples for body parts in three ways: 1) Starting with the PS solution, take the part locations and orientations and draw a shape at random from a prior over part shapes; 2) Draw a random body from our DS prior, 3) Draw parts independently from a prior over locations, orientations, and shapes. These independent part priors are learned from an annotated training set of DS models (see next section).

A first iteration of BP estimates the location of the torso. The appearance color model is then built as a histogram in CIE a*b* space from the pixels that correspond to the head and upper part of the torso. During each subsequent iteration of BP, samples are generated at each node (part) by a random walk from the current samples or proposed by the neighbors: for example, given a likely torso and a likely lower arm, a new upper arm is sampled conditioned on the parent shape and the child location. The conditioned proposal exploits the part length parameter in the model formulation (5) to generate a sample that is likely to connect the torso and the lower arm. For each new sample, we evaluate its acceptance probability as in PBP [17], with the difference that good samples are never removed from the node.

Since the DS model is learned from SCAPE, the joint positions of neighboring parts overlap exactly. This means the learned model variance in the springs is zero, which makes inference difficult. Consequently, for inference, we artificially inflate the variance for the springs connecting parts to give non-zero probability to configurations of parts that are not connected. This creates a *loosely connected* DS model.

7. Experiments

Our central hypothesis is that a more accurate model of body shape should result in a more discriminative likelihood model and consequently more precise estimation of body pose in images. To test this hypothesis, we perform several experiments and compare our results with the published state of the art. Experiments are run on the Buffy dataset as used in Sapp *et al*. [30]. To learn the independent part priors used in sampling, we annotated images in a Buffy training dataset, not present in the test set, using the annotation tool described in Section 4. The features for the likelihood are learned from the training set of images in Section 4, none of which are from the Buffy series. The scalar parameters in the likelihood are estimated from the Buffy training set using the method in [26].

The Buffy dataset comes with "stickmen" annotations, defining the ground truth end points for each body part. Error is computed as the Percentage of Correctly estimated body Parts (PCP). A body part is correctly estimated if its end points lie within half of the ground truth segment's length from the ground truth end points [10]. The stickmen ground truth data has joints in different locations than the DS body model. Consequently we estimate a correction term mapping the DS joint points to the stickmen joints using annotated images from the Buffy training set. This correction term is simply the mean 2D offset of the joint location in the coordinate system of each part.

We test 3 different versions of our model: 1) The full model but with a *uniform likelihood* function (NL). This tests that our inference method is actually exploiting the likelihood and is not just based on the good initial proposals. 2) A model with no shape variation (NS). This uses the *mean shape* of the DS puppet and lets us tease apart the effect of our likelihood function from the shape model. 3) The full DS model (DS), with 4 shape coefficients per part.

Since our inference relies on a PS model [3] for initialization, we take this to be the baseline. A more discriminative likelihood should improve performance. A fully fair comparison with the PS baseline is difficult. It is not possible to simply interchange the three key components: inference, model (prior), and likelihood. They interact in ways that make separate analysis difficult. The PS model in [3], for example, uses a discretized state space and optimal inference. Our method is at a disadvantage in that the inference is stochastic.

Performance results are reported in Table 7. In our inference framework the DS model effectively refines the results of the baseline method. Additionally the value of the shape space is seen in the significant improvement over the mean shape model (NS). In Table 7 we also report results from [1], which is not an articulated model, but represents parts with mixtures of non-oriented pictorial structures.

Figure 8 shows several representative examples where the DS model (solid red) improves the baseline PS result (dashed green). The performance of the inference is dependent on the performance of the baseline PS model in providing a good initialization, and on the correctness of the scale estimation. Figure 9 shows some representative examples of failures.

Our results are essentially the same as the CPS method

Figure 8. Estimated body pose – examples where the DS model improves on the PS baseline. DS is solid red, PS is dashed green.

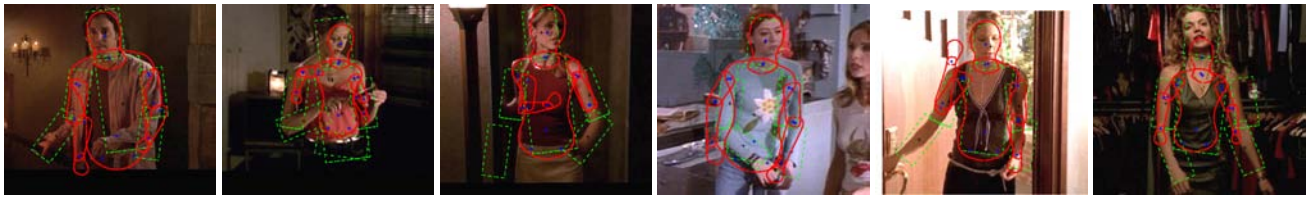

Figure 9. Estimated body pose – representative failure cases. DS is solid red, PS is dashed green.

Method	Torso	Head	U. Arms	L. Arms	Total
Baseline (PS)	97.0	92.3	86.3	52.1	77.7
Our (NL)	99.2	97.9	90.9	10.4	66.6
Our (NS)	99.2	97.5	94.0	50.4	80.9
Our (DS)	99.6	99.2	94.7	62.8	85.6
Eichner *et al.*	98.7	97.9	82.8	59.8	80.1
CPS	100	96.2	95.3	63.0	85.5
Yang *et al.*	100	99.6	96.6	70.9	89.1

Table 1. PCP scores (see text) for our model without likelihood (NL), our model with a fixed shape (NS), and our full model (DS), with shape variation. PS is the implementation of [3]. We also compare with the current state of the art: CPS [30] and Yang *et al*. [1].

[30] which uses a more a sophisticated cascade search method and richer likelihood models that span multiple parts. Both cascaded search and the extended likelihoods are well suited to the DS shape model, suggesting room for further improvement.

8. Conclusions

Deformable structures are a generative model of 2D human shape in images and define a prior probability of body shape and pose. The DS model is learned from 3D bodies covering a wide range of poses and camera views. While the formulation is simple, using Gaussian potentials, it is expressive enough to capture significant shape deformations. It has the benefits of pictorial structures and yet also models pose-dependent body shape. What differentiates the DS model from other 2D shape representations is the fact that the shape of a person is expressed as a factored probability over parts. We exploit the shape representation to learn a likelihood model that takes into account the part contours. To do so, we use the DS model as a novel puppet for annotating training imagery. We show improved human pose estimation over the basic pictorial structures model and essentially the same accuracy as the CPS method which uses

more sophisticated inference and likelihood models [30].

In future work we will expand the range of camera views and include separate DS models for side views which we did not consider here. Additionally, the SCAPE model represents a wide range of body shapes. We will consider DS models with individual body shape variation. This could involve local shape terms in the parts or a loopy model with a global shape parameterization. Finally, we have seen that the DS model captures body shapes in real images but this might be improved by adding a clothing model [16].

Acknowledgments. We thank M. Andriluka and B. Sapp for making their code and dataset available. This work was supported in part by a grant from the NIH-NINDS EUREKA program (R01–NS066311).

References

[1] Y. Yang, and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *CVPR*, pp. 1385–1392, 2011.

[2] J. Charles, and M. Everingham. Learning shape models for monocular human pose estimation. *ICCV Workshops*, pp. 1202–1208, 2011.

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR*, pp. 1014–1021, 2009.

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of PEople. *SIGGRAPH*, 24(3):408–416, 2005.

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. *ICCV*, pp. 1365–1372, 2009.

[6] M. Bray, P. Kohli, and P. Torr. PoseCut: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. *ECCV*, pp. 642–655, 2006.

[7] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models – Their training and application. *CVIU*, 61:38–59, Jan. 1995.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, pp. 886–893, 2005.

[9] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. *ECCV*, pp. 228–242, 2010.

[10] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. *BMVC*, 2009.

[11] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. *CVPR*, pp. 66–73, 2000.

[12] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[13] V. Ferrari, M. Marin-Jiminez, , and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR*, pp. 1–8, 2008.

[14] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Computers*, 22(1):67–92, Jan. 1973.

[15] O. Freifeld, A. Weiss, S. Zuffi, and M. Black. Contour people: A parametrized model of 2D articulated human shape. *CVPR*, pp. 639–646, 2010.

[16] P. Guan, O. Freifeld, and M. Black. A 2D human body model dressed in eigen clothing. *ECCV*, pp. 285–298, 2010.

[17] A. Ihler and D. McAllester. Particle belief propagation. *AISTATS*, pp. 256–263, 2009.

[18] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. *CVPR (2)*, pp. 180–185, 2001.

[19] M. Isard. PAMPAS: Real-valued graphical models for computer vision. *CVPR*, pp. 613–620, 2003.

[20] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. *BMVC*, pp. 12.1–11, 2010.

[21] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. *CVPR*, pp. 1465–1472, 2011.

[22] M. Kumar, P. Torr, and A. Zisserman. OBJ CUT. *CVPR*, vol. 1, pp. 18–25, 2005.

[23] C. Liu, L. Sharan, E. Adelson, and R. Rosenholtz. Exploring features in a Bayesian framework for material recognition. *CVPR*, pp. 239–246, 2010.

[24] D. Parikh and L. Zitnick. Finding the weakest link in person detectors. *CVPR*, pp. 1425–1432, 2001.

[25] L. Pishchulin, A. Jain, C. Wojek, T. Thormaehlen, and B. Schiele. Good shape: Robust people detection based on appearance and shape. *BMVC*, pp. 5.1–5.12. 2011.

[26] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[27] D. Ramanan. Learning to parse images of articulated bodies. *NIPS*, pp. 1129–1136, 2006.

[28] D. Ramanan and C. Sminchisescu. Training deformable models for localization. *CVPR*, vol. 1, pp. 206–213, 2006.

[29] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. *CVPR*, pp. 422–429, 2010.

[30] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. *ECCV*, pp. 406–420, 2010.

[31] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. *IJCV*, 54(1-3):183–209, 2003.

[32] L. Sigal and M. Black. Predicting 3D people from 2D pictures. *Proc. IV Conf. on Articulated Motion and Deformable Objects (AMDO)*, pp. 185–195, 2006.

[33] L. Sigal, M. Isard, H. Haussecker, and M. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15-48, 2011.

[34] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. *CVPR*, pp. 605–612, 2003.

[35] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. *ICCV*, pp. 723–730, 2011.

[36] H. Wang and D. Koller. Multi-level inference by relaxed dual decomposition for human pose segmentation. *CVPR*, pp. 2433–2440, 2011.

[37] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. *ECCV*, vol. 3, pp. 710–724, 2008.

[38] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. *CVPR*, pp. 1705–1712, 2011.